第 3 章

数字化的行为

3.1 消费者数据来源

3.1.1 数据收集

在将目光投向数字化的消费者行为数据前，我们首先需要了解传统的市场调研方法：主要包括**观察法**、**调查法**和**实验法**。这可以帮助我们回答一些逻辑上的问题：我们到底需要什么样的数据，以及如何获取我们需要的数据。

1. 观察法

观察法是指通过观察相关人员、行为和情景来收集原始数据。调研人员可以通过观察消费者的行为来探寻一些无法从询问消费者当中获得的信息。人种学研究就是其中的方法之一，该调研方法在自然状态下观察消费者并与他们互动。观察者可能是训练有素的人类学家、心理学家，或者是公司的调研人员和经理。假设一个厨房用品企业希望通过人种学研究来为它们的产品优化提供新的视角，那么它们可以真实地走进厨房，观察使用者如何使用设备：使用多大分量的食材，家庭成员如何在厨房中进行互动，如何在不同的厨具之间切换以及其他一些使用上的细节。

2. 调查法

调查法是收集数据最常用的方法，最适用于收集描述性数据，可以帮助企业了解人们的认知、态度、偏好或购买行为。其主要优点在于其灵活性，可以在许多不同情况下获得不同的信息。但获得的数据质量无法保证：被调查者有时可能出于隐私、时间等原因而拒绝回答一些问题，抑或是出于表现自己或者帮助调查者的原因，给出不准确的答案。此外，调查法往往受时间限制，无法得出有效的因果结论：相同时间所获得的信息无法为结论提供因果效力。因此，调查法往往只用于描述性数据的收集，无法提供对产品的深入洞察意见。

3. 实验法

实验法选择配对的实验组和控制组，从而考察组间被试的反应差异。实验法适用于收集反映因果关系的信息，因此被广泛应用于市场营销和消费者行为领域的研究中。其劣势在于，完美的实验条件是很难达到的。首先，在实验过程中，许多无关的影响因素是无法排除的，因此实验效果中往往混杂着非实验因素的影响结果；其次，实验对象和实验环境的选择往往有一定的限制，因此结果很难具有充分的代表性，其结论也带有一

定的特殊性，其应用范围往往是很有限的；最后，实验法对调查者的能力要求比较高，花费的时间也往往比较长。

上述这些市场调研信息可以通过邮件、电话、个人访谈及网络等方式获得,其中网上营销调研是目前广泛使用的一种适用于定量调研的方式。企业可以通过发布问卷、发送电子邮件等方式邀请消费者回答。一些在线的问卷平台也支持网上的实验设计，方便调研人员观察不同的价格、标题和产品属性对营销措施的影响。

网上营销调研的速度快，成本低。主流的问卷平台，如国内的问卷星、见数（Credamo），国外的 Mturk 等平台，拥有庞大的用户群体，这使得收集 2000 人左右的样本只需要花费一个晚上的时间就可以完成。同时，这一方式能够更准确地寻找到目标人群。虽然增加人群的限制条件需要给平台支付额外的人均费用，但相较于传统的调研方式，借助调研平台的网上营销调研免去了企业自行寻找目标群体的成本，也很好地避免了选择到错误的目标调研群体或是没有完全覆盖目标调研群体的风险。

3.1.2 在线数据

企业的在线数据通常通过与消费者的交互获得，这些数据可以来自在平台上广告的推送、自有网站或 App、事件监测（埋点）和小程序端等渠道。对于企业来说，这部分数据易于获取，也反映了消费者与品牌或产品的直接接触。下面将根据不同的数据来源对企业的在线数据进行介绍。

首先我们来关注来自广告推送的消费者行为数据。如今，我们时常会收到许多类型的广告推送：想象这样一个情景，假设你从床上醒来，打开手机想要看看今天的天气。此时天气软件的首页就向你推送了某雨伞品牌官网的购买界面，你饶有兴趣，点击并跳转到了这一界面。此时，你的观看广告和点击链接行为就通过广告推送被记录了。由此获得的数据体量巨大，如果企业能进一步获得具体的个人信息，那么这类数据的价值将不容小觑。但在现实中，推送平台或应用出于保护自身的利益以及遵守隐私协议等原因，并不会提供具体到个人的信息，因此来自广告推送的消费者行为数据的价值受到限制。

在这方面，从自有网站上获取的消费者行为数据就优于来自广告推送的消费者行为数据。企业可以通过客户端脚本检测（针对网页）和软件开发工具包（针对应用）等方法，记录下需要检测的页面上每次访问的相关数据，如访问时间、访问来源、停留时间等。

这类数据的获取技术成熟，在使用权上不存在争议，因此应用也更加方便。同时，这类数据相较于广告推送的观看或点击，更加接近“购买行为”，因此有着更高的价值。具体来说，这些数据所面对的访问网站或使用应用的消费者，他们本身对产品是有一定的兴趣，因此其价值也就更高。比如，一个点击链接进入毛巾商品购买界面的消费者，显然会比在其他平台被动观看毛巾广告的消费者更有可能购买这一产品。

然而，这些数据的获取方法存在一定局限性：出于技术层面的原因，对于网站端数据，客户端脚本检测方法所获得的数据不够准确，也无法记录消费者在网页上进行的更加复杂的交互，如点击播放视频、调整商品选项等；对于应用端的数据，类似方法的在

技术上的实现难度也较高。

事件监测（埋点）就能够很好地帮助企业完善来自网站和应用上的消费者行为数据。事件监测（埋点）在基础的监测程序上关注特定的用户行为或事件，并在其发生时进行处理和记录。对于网站来说，事件监测（埋点）关注的事件就是在某一网页上发生的交互，如点击播放商品展示视频。

事件监测（埋点）方式能够适应企业获取复杂程度更高的交互数据的需求，有着更广泛的适用性。目前一些监测工具已经实现了操作更加简便的可视化埋点以及抓取所有交互的无埋点（全埋点）功能，大大方便了企业进行事件监测的操作。

随着社交媒体的兴起，从社交媒体（如微信）提供的公众号、小程序中获取消费者行为数据也成为一种常见的方式。公众号与小程序可以被视作是一种特殊的网站，企业可以使用一些监测脚本，如腾讯公司官方提供的腾讯移动分析（Mobile Tencent Analytics，MTA），从中获取消费者行为数据。同样，从这一渠道获取的数据可以使用事件监测（埋点）方式进行完善，但在具体的技术层面有着些许差异。

3.1.3　购买第三方数据

从自有渠道获取的消费者行为数据固然便利，但海量的第三方数据蕴含的商业价值不容忽视。事实上，今天的企业常常会为使用外部数据投入大量的资金，第三方数据可以通过一些来自社交媒体、电商平台、广告行业、政府部门以及专门的数据交易平台来进行交易。

不同渠道提供的行为数据有着不同的优势和不同的获取手段。社交平台（如微博）的海量用户意味着高水平数据的用户多样性，能够在进行市场调研时覆盖更为广泛的受众群体，收集到更具代表性的数据。电商平台渠道则意味着与具体业务的密切关联。比如，阿里旗下生意参谋网站（https://sycm.taobao.com）就能够为店家提供页面浏览量（page view，PV）、支付金额、浏览来源、访客位置等实时信息，帮助电商商家进行实时监控与调整。广告行业往往积累了许多关于广告投放效果与用户交互的数据，因此其提供的数据能够提供有关于广告效果评估、投入估算等方面的有价值的信息。最为特殊的是政府来源的第三方数据，该类型的数据往往能够涵盖广阔的领域。比如，国家统计局城市社会经济调查司主办的《中国城市统计年鉴》反映了从 1985 年开始，全国 656 个建制城市的城市分布情况、人口、劳动力及土地资源、综合经济、工业、交通运输、邮电通信、贸易、外经、固定资产投资、教育等方面的数据。如此广泛的覆盖范围和类别决定了政府来源的数据的高价值，但前提是应用者必须掌握一定的利用手段，能够筛选出其中有效的信息。政府来源的数据可以被广泛地利用于各个与营销相关领域，如果需要使用类似的政府数据，使用者可以通过向国家统计局提出申请，或直接通过官方渠道购买获得。

但随着互联网用户隐私保护概念的强化，第三方数据的**精细程度正在降低**。很多第三方数据服务商现在提供的是“打包”过的数据，即反映人群中个体拥有某一共同属性概率的数据。比如，在 1000 人的数据中发现有 800 人有可乐的消费记录，那么“打包”

处理后的人群数据就会这样报告：这一人群中有 80%的个体有可能消费可乐。这样的数据能够较好地规避隐私安全方面的担忧，其质量也可以通过第三方的数据机构来进行验证。

此外，企业只能应用这些数据，却无法“持有”这些数据，也无法保证数据的使用效果。比如，一个经营矿泉水业务的企业购买了某一区域 1000 个被第三方标记为“矿泉水”消费者的数据，并通过该平台的推送系统向这 1000 名消费者发布了该企业的矿泉水广告。整个过程中，企业并不“持有”这些数据，自然无法保证这 1000 名消费者中，到底有多少人真正是企业的目标客户，这样的推广质量显然是无法得到保障的。因此在第三方数据的购买上，企业更多地愿意选择大体量第三方数据供应者，如腾讯、阿里巴巴等。

3.2 消费行为数据类型

3.2.1 常见数据类型

提起我们生活中常见的营销相关数据，我们首先可能想到的就是交易中产生的销售记录、发票、运费等数据来源提供的数据，如销售量、交易时间和交易来源等。此类拥有高度统一格式且能够被轻松整合的数据叫作**结构化数据**。结构化数据在营销实践中被广泛应用，它不仅可以是与销售相关的日期、产品、客户信息，也可以是来自外部的其他可能相关的信息，如前文提到的国家发布的空气质量数据等公开数据。

企业、研究者已经使用结构化数据指导了很多成功的营销实践。例如，企业的营销人员可以通过电商平台的程序，提取其在线专卖店某项产品销售记录中的用户收货地址，制作出该产品的在不同区域的销售分布图，进而制定更有区域针对性的营销方案。

结构化数据一个显而易见的优势是便利性。一方面，结构化数据有很高的拓展性。结构化数据的相关技术已经十分成熟，企业有着多样的手段对其进行处理，不仅可以通过外部数据来进行辅助的分析，也可以轻松地将其应用到机器学习等新兴的分析工具上。另一方面，结构化数据的使用成本更低，更受中小型企业的青睐。结构化数据将数字、符号、名称等信息使用统一的结构（如电子表格）表示，这使得在收集过程中，企业通过较低的人力、技术成本就可以收集到大量的结构化数据并投入使用。

高度统一的形式也带来了结构化数据的一个劣势：其商业价值的可挖掘程度有限。结构化数据的形式决定了它在内容的复杂程度上是有限的，一个 1000 人的结构化数据表格的大小甚至不会超过 500 KB。有限的信息量决定了其相对固定的用途。例如，单纯的季度销售量数据只能反映用户在一段时间内的购买行为，而无法从中得知用户具体的心理过程，也就无法得知真正影响顾客决策的产品、环境和促销等方面的因素。与内容更加复杂、全面，用途更加多样化的非结构化数据相比，相同体量的结构化数据在商业上应用价值的可挖掘性不高。此外，结构化数据也并不总是有效。现实中，消费者的行为受很多因素影响，我们很难从简单的数据中得出有效的预测模型。这使得很多营销人员在使用统计方法对内容上有限的结构化数据进行解读的过程中，不可避免地会产生

一些有局限的结论。事实上，很多企业在使用结构化数据的分析结果时，得出的正确结论并不能很好地指导具体的实践。正因如此，企业和研究者现在越来越关注包含更多细节、更加全面的非结构化数据。

非结构化数据就是除了结构化数据之外的所有数据。非结构化数据的形式可能并不符合大多数人对于数据的直观想象：它的形式不定，它可以是文本，也可以是非文本。例如，用户在社交媒体上发布的产品相关的文字、图片，甚至是视频，这些形式各异的信息都属于非结构化数据。

非结构化数据蕴含着可以挖掘的大量有价值的信息，营销人员可以使用非结构化数据中的细节更好地了解消费者，制定有效的营销策略。但形式不统一的非结构化数据往往需要通过一定的处理方式转化，才能应用到营销实践中。这意味着非结构化数据相较于结构化数据并不容易使用。不过随着人工智能技术在非结构化数据使用上的日益成熟，这一差距越来越小。目前也涌现出了许多能够帮助企业处理非结构化数据的在线平台，如之前提及的见数等平台。接下来，我们将对几种不同的非结构化数据进行介绍。

3.2.2　文本数据

文本数据是一种典型的非结构化数据，其内容可以是字母、汉字、非数值的数字及其他符号。常见的文本数据包括电商平台上的用户评论、社交平台上的用户生成内容（如微博、朋友圈等）、第三方平台上的产品评价（如小红书的产品分享等）或共同协作内容（如维基百科，百度百科等）。

文本数据的优势在于它包含了大量潜在的有价值信息，这是因为消费者和潜在客户产生的相关文本内容往往反映了他们对品牌、产品的态度以及其他一些可能与用户进行决策时心理过程有关的信息。通过对这些数据进行定量、定性的分析，企业可以对比消费者评价与企业期望的差距，从而对营销方案做出针对性调整。现在大量的企业正在投入大量资金使用舆情监测工具来完成实时的舆情监测，这一点也从侧面反映了在实践中企业对于文本数据的重视。虽然文本数据潜力巨大，但是它的使用有一定的技术要求：文本数据长度不统一，形式不相似，营销人员需要在实践中使用文本挖掘技术。在营销领域，常见的文本挖掘主要使用词频转化、文本向量化的方法，而向量化后的文本数据可以进一步使用自然语言处理（natural language processing，NLP）等人工智能方法处理。

词频转化通常是通过对特定文本中某个词或词组的频率计数进行的。这种方法可以将非结构化的文本数据压缩并结构化，也就是将原本定性的文本数据转化为定量的词频数。在此类分析过程中，我们需要事先将文本在结构上分解出对象（产品、服务或活动等）和观点部分，并使用词库对观点部分的内容进行处理。以观点挖掘为例：对消费者的产品评价这样的文本数据进行的文本挖掘，可以通过统计文本中代表正面评价（好用、实惠等）和负面评价（劣质、不值等）的词语出现的总数，得出一个数值化的消费者总体态度。通过这样的方法，不仅可以推断消费者的情绪（情感）、态度，还可以通过特定词库就他们的人格特质展开分析。

对于词频转化来说，一个成熟的词库必不可少。在相对比较成熟的领域（如情感分析），我们可以很容易地在网络上找到共享的优质词库，或是直接通过在线的专业分析平台使用它们自己构建的词库。但当我们对文本数据的分析聚焦在某个不是那么热门的领域时，我们可能就需要自己构建词库。此外，反讽句子在近几年的热点事件中越来越常见，因此在对热点进行内容分析时，我们可以考虑辅助使用反讽词语的词库。比如，“呵呵”“卧龙凤雏”等表达在文本中可能不再代表原有的积极意义，反而可能是“阴阳怪气”，可能会对分析的结果产生很大的不利影响。近年来已经有许多研究利用国外社交媒体平台上带有反讽标签（如#irony、#sarcasm）的文本来构建反讽表达的词库，也有少量的中文研究构建了此类型的词库，我们在词频转化过程中可以考虑将这些词库纳入分析过程。

文本向量化方法是另外一种文本挖掘的方法，它将文本数据组织成一个文档特征矩阵，矩阵中的每一行代表一个文档，每一列代表一个选定的特征词。文本向量化方法尤其适合大规模的文本数据，因为这样的处理格式化了文本数据，方便了之后适合处理大量文本数据的人工智能方法的使用。

文本数据被越来越广泛地使用，其中一个热点就是人工智能文本分析处理技术：自然语言处理。自然语言处理是人工智能领域中的一个焦点技术。这一技术聚焦于使用自然语言进行人机沟通。自然语言处理技术使对大量文本数据进行智能化的处理成为可能。市场营销中的自然语言处理模型应用在 10 年前就已经出现，但大多数研究和实践都是最近才开始的。自然语言处理模型的应用涵盖了四个主要的实质性领域：P2P 平台，社交媒体，在线的消费者评论、企业公告和广告，网站和搜索结果。自然语言处理模型可以应用的文本类型广泛，如新产品公告、社交媒体聊天、产品评论、客户反馈、网页内容、客服互动等文字内容都有着对应现成的模型。对于不同类型的文本数据，自然语言处理技术可以提供不同的分析功能。比如，针对一家饭馆的在某平台上的在线点评，可以通过自然语言处理技术分析不同用户的评论内容中所包含的情绪及情绪效价，从而进一步为针对在线口碑问题的解决方案提供建议。针对自然语言处理技术的介绍我们将在第 4 章中进行更详细的论述。

文本数据的应用过程一般包括数据预处理、文本信息提取、选择常用的文本分析指标。在任何数据分析之前，都要先将文本数据预先“清洗”处理，进而产生类似 excel 表的干净的数据。常用的工具有 R 语言和 Python 语言，两种编程语言都有一套易用的数据预处理包。

随着技术的发展，文本数据的挖掘深度正在不断地提高。在个体层面上，情感和满意度是过去最常用的测量变量，最近几年不断有研究拓展可以从文本数据预测的变量，如从文本数据中提取的语言的真实性和情绪性，甚至也有更加复杂、高级的心理学测量变量，如性格类型和建构水平。相信将来一定会有越来越多潜在的可以借鉴应用到消费者相关文本数据分析的变量。

3.2.3 视频流数据

视频流数据在市场营销领域近几年的一个热门应用是基于增强现实（augmented

reality，AR）的营销。增强现实技术是一种将计算机生成的信息投射到现实的技术，这一技术的前提就是视频流技术所带来的低延迟视频数据传输。而增强现实营销是指将增强现实体验与其他媒体渠道整合，从而创造品牌与消费者价值，实现营销目标的一种营销策略。

增强现实技术可以为消费者提供独特的身临其境的数字体验，使品牌进行更具有互动性的营销实践并获得更加贴近消费者实际生活的设计洞见。例如，在 2020 年 10 月，相机企业 Snap 在伦敦的卡尔纳比（Carnaby）街揭幕了“城市画家”活动。利用增强现实工具，用户可以在街道的商店上方虚拟地“喷涂油漆”，并使用用户预先创建的壁画对其进行装饰。Snap 公司希望通过绘制主要地标来创建一个共享的虚拟世界，带给消费者崭新的使用体验。在一段宣传视频中，该公司解释说，通过“使用各种数据源、360 度图像和社区快照——我们能够建立实体世界的数字表示形式”。这一活动也为疫情期间的旅游业带来了新的灵感。随着增强现实技术的发展，毫无疑问，将会有更多的品牌参与其中。

3.2.4　互联网行为数据

请想象自己正在浏览淘宝网，选购自己心仪的衣物：首先，你会进行商品的搜索；然后选中了一项符合你要求的商品，并点开商品详情页进行浏览；你在详情页放大了商品缩略图，阅读了商品详情页中的原材料信息；最后你决定将这一商品加入购物车或选择立即购买。整个过程都会被企业记录，形成互联网行为数据，用来指导企业未来的营销实践。互联网行为数据就是能够广泛反映用户在互联网所进行行为的数据，与之相对的概念是用户数据（用户标签及个体信息）和业务数据（如购买记录）。具体的概念框架如图 3-1 所示。

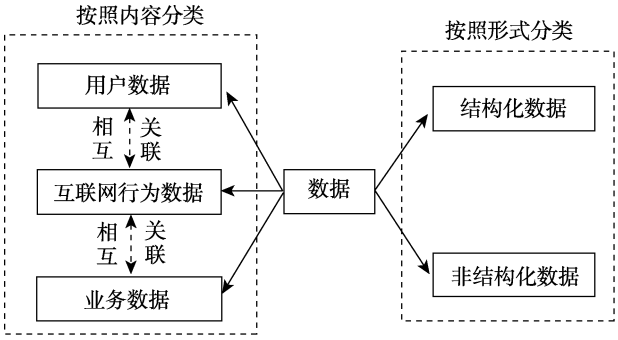

图 3-1　数据的分类

互联网行为数据主要反映四个方面的信息：访问频率（页面的访问次数、用户的访问频繁程度和深度等）、行为转化率（有多少用户加入购物车、形成订单或完成付款）、用户的访问时长以及用户的访问质量（有多少用户短暂浏览后就退出）。表 3-1 就是一份来自阿里云的互联网行为数据其中的一部分及其概述，表 3-2 为行为与商品时间数据。

表 3-1　阿里云的互联网行为数据

列名称	说明
用户 ID	integer
商品 ID	integer
商品类目 ID	integer
行为类型	string，枚举类型。其中 pv 等价于点击商品页面，buy 等价于商品购买，cart 等价于将商品加入购物车，fav 等价于收藏商品
时间戳	行为发生的时间戳

表 3-2　行为与商品时间数据

时间戳	商品 ID	商品类目 ID	行为类型	用户 ID
12308	4961876	2837163	pv	1512057220
12308	4271826	1449177	fav	1512057285
12308	4291446	1609467	buy	1512116585
12308	2428637	1464116	pv	1512116614
12308	4291446	1609467	pv	1512116655

互联网行为数据连通用户数据和业务数据，构成了完整的事件过程，这为企业的营销实践提供了更多广阔、深入的洞见。对于设计团队来说，互联网行为数据可以帮助他们提升用户体验、优化界面与功能；对于运营团队来说，互联网行为数据可以帮助他们更加精准地投放文案、活动、广告；对于公司来说，互联网行为数据可以帮助它们更深入地了解业务，以更快地对市场反应做出应对。常用的互联网行为数据分析方法包括事件分析、分布分析、留存分析、漏斗分析及路径分析。关于这几种分析方法，我们将在第 5 章中详细展开，论述从数据中提炼消费者行为的逻辑。

针对互联网行为数据的利用，有许多公司已经有了相当成功的实践。eBay 于 2016 年成立了大数据分析平台，对海量的消费者行为数据进行分类，并改进了其网页设计的流程。具体来说，eBay 的设计流程从由设计师主导变为由行为数据主导。基于此努力，eBay 不仅极大地优化了自身的业务过程，吸引了海量的新用户，也实现了可观的利润增长。

3.3　大数据存储

3.3.1　数据规模

按照国际商业机器公司（International Business Machines，IBM）与甲骨文（Oracle）公司对大数据的定义，大数据的基本特征可以用五个 V 来总结（volume、variety、velocity、value 和 veracity），即体量大、多样化、速度快、价值密度低、真实性。体量大意味着需要处理海量、低密度的非结构化数据。在实际应用中，大数据的数据量通常高达数十太字节（terabyte，TB），甚至数百拍字节（petabytes，PB）。多样化指的则是众多形式的数据类型，如文本数据、图片数据、音频数据甚至是视频流数据。速度快意味着大数

据的接收与处理几乎都是实时或零延迟运行。截至 2022 年，中国网民数量已超 10 亿，每周网络时长将近 30 小时，如此庞大的日常活跃用户数量将会在极短时间内产生大量的数据。价值密度低和真实性在某种程度上是互通的：大数据虽然蕴含着巨大的商业价值，却是建立在合理的发掘方法以及可靠的信息来源上的。一方面，时长为 10 分钟、大小为几百兆的视频文件，可能具有价值的部分仅仅只有几秒——如此低密度的有价值信息显然需要一定的工具来帮助挖掘；另一方面，现在存在着许多网络水军或是机器人团队生成的数据，这些数据往往并不能反映客观的用户行为，甚至会污染已有的数据库。这些性质共同挑战了传统的数据存储方式。

面对来自全世界不同领域、不同形式的大数据的爆炸式增长，除了海量的数据规模，数据本身异构型、应用所要求的高时效性的需求不仅仅意味着存储容量的压力，还给系统的存储性能、数据管理乃至大数据的应用等方面都带来了严峻的挑战。举一个交通领域的例子，北京市交通领域的工作者一直致力于提升交通智能化水平，为市民提供更加安全、高效、舒适的出行服务。为了实现这一目标，北京市交通运行智能化分析平台采用了全新的数据库技术来存储数据。该平台的数据来源非常丰富，既包括传统的交通工具，如公交、轨道交通、出租车等，又包括新兴的共享出行方式，如网约车、共享单车等。这些数据从不同的角度反映了交通出行的状况，为交通智能化分析提供了有力的支持。此外，部分数据还源于问卷调查和地理信息系统，如车辆每天产生的出行轨迹，交通卡的记录，手机的 GPS 定位轨迹，出租车运营公里数，每个停车场的电子停车收费系统数据，定期上门调查所在地家庭，等等。这些数据不仅在体量和速度上都达到了大数据的规模，也涵盖了文本、音频、图片、视频、模拟信号等不同的类型。

3.3.2　SQL 本地数据仓库

20 世纪 70 年代，如今最为人所熟知的关系数据模型的概念被提出，由此奠定了关系型数据库的基础。关系数据模型由多个二维表以及表之间的逻辑关系组成：想象一所大学正在创建一个关系型数据库，以对自己学校的教学数据进行存储和管理，那么这个模型可能会包含教师信息表、课程信息表、教师课程表，并将各个表通过其中某一共有列的元素相互关联——如教师信息表和教师课程表中共有的教师工号，课程信息表和教师课程表中共有的课程编号。随之出现的结构化查询语言（structured query language，SQL）技术，里程碑式地实现了关系数据模型在计算机上的实际应用。SQL 技术是一种应用于关系型数据库的标准计算机语言，它可以帮助用户对目标数据进行定义、处理、查询。

SQL 技术有着一些显著的优势。首先，SQL 通过形式上极其标准的二维表以及表链接规范了数据库中所有的数据。在处理结构化数据时（如教学信息数据），显然 SQL 是最为高效的方式。其次，SQL 对用户来说是相当友好的一门语言。它的逻辑与现实逻辑几乎一致，基本不需要编码，使用和维护都极其方便，并且已经存在许多成功的实践可供参考。最后，运用 SQL 技术构建的本地数据仓库在安全性能上也会高于本节后面将提及的其他几种新型数据存储方式。

进入大数据时代，曾经风靡一时的 SQL 技术在面对海量的数据时往往捉襟见肘。如前文所述，无论是网页还是应用，现如今的数据体量绝非几十年前可比。如果依然使用硬盘读写这些数据，存储、管理动辄数百 PB 的数据是不可能的任务，不同二维表之间复杂的逻辑关系也会导致其读写速度的严重下滑，并且 SQL 的横向（几种数据）、纵向（多少数据）拓展也是一项十分艰巨的挑战。SQL 数据库纵向的拓展只能通过提升现有服务器或直接更换更加高性能（内存、CPU 等）的服务器才能实现，这样的升级会消耗大量的人力、物力和财力。更糟糕的是，这种拓展往往吃力不讨好——硬件的更新换代与技术的需求都在快速地发生，与之赛跑显然并不是明智之举。而横向的拓展所耗费的精力也同样巨大且是一个无底洞：完美的、包容一切所需的数据框架在这个信息爆炸的时代是只能存在于理论上，而拓展过程中所需开发人员的人力成本却是实打实的。大数据时代数据的多样性，对于 SQL 技术而言也是难以突破的一个瓶颈。二维表结构相当于早期简单且结构化的数据，但对于本章之前提及的非结构化数据，以关系数据模型为基础的 SQL 技术就束手无策了。这也是 Web 2.0 时代传统 SQL 本地数据仓库没落的原因之一。

需求在一定程度上能够促进技术的发展，在面对存储、管理海量数据的需求时，NoSQL（not only SQL）技术脱颖而出。NoSQL 是所有的不同于传统关系型数据库的数据库管理系统。常见的 NoSQL 应用有 MongoDB、Redis、CouchDB 等，虽然有所差异，但这些 NoSQL 应用基本都通过牺牲固定的数据结构，换来了许多适应数据新特性的功能。一是卓越的读写性能：由于无需解析 SQL 层，读写操作的速度大大提升。二是在数据库拓展方面的便利：NoSQL 技术允许用户轻易地链接不同类型的数据库，也可以被安置在不同的服务器（分布式系统）上。这使得数据库的横向与纵向拓展都具备着相当程度的敏捷性。

尽管 NoSQL 技术更加适应当下的环境，但今天我们仍然在某些领域大量使用着 SQL 技术。因为 SQL 技术有着其不可替代的特性：稳定、安全、便利。对于数据横向、纵向规模增长速度相对稳定的行业，SQL 在拓展方面的缺陷是可以接受的；SQL 不需联网，在本地的使用也不用考虑分布式系统带来的安全隐患，因此在安全方面具备着压倒性的优势；SQL 易于操作，不需要具备过硬的计算机知识即可建立相关的数据库。符合这 3 个需求的行业往往依旧选择传统的 SQL 技术构建数据库，一个典型的例子就是银行业：相对稳定的业务（横向规模），可以预测的用户增长速度（纵向规模），高度的安全需要以及简单可行的操作。

3.3.3 第三方云存储

分布式系统能够带来众多节点之间的通信成本上升，在不同的服务器之间存储数据会对安全性和保密性造成可观的威胁。随之出现的云计算技术为今天的大数据技术领域带来了深远的影响。

云存储是一种云计算模型，该模型可以通过云计算服务提供商将数据和文件存储在互联网上，而用户可以通过公共网络或专用网络连接访问这些数据和文件。提供商安全

地存储、管理并维护存储服务器、基础设施和网络，以确保用户在需要时能够以几乎无限的规模和弹性容量访问数据。

一般的云存储的类型分别为对象存储、文件存储和块存储。虽然不同的云计算服务提供商对于其提供的云存储类型有着不同的称呼与描述，但在定义上是类似的。对象存储允许用户在不使用文件系统的情况下远程存储和检索对象存储中的非结构化数据。存储项在云中只是一个抽象“对象”，这意味着应用开发人员可以保持尽可能高的灵活性，在云中拥有一个基本上不限量的自由格式的数据存储，同时只需按存储量和传输量付费。高自由度的对象存储技术主要应用在非结构化数据的存储（如针对网站的备份记录）。而文件存储为在云服务器上运行的操作系统进行文件共享，然后操作系统再将该文件提供给虚拟机中运行的应用。这意味着，应用迁移到云端后仍可继续使用一直在使用的文件存储，并且用户可以根据需求增长进行扩容。文件存储形式是在生活中相当常见的一种云存储类型，许多企业的文件共享就属于此类。块存储类似于文件存储，区别在于其更佳的性能以及更高的相关硬件要求。块存储的一个主要应用就是联机事务处理（online transaction processing，OLTP）系统。该技术可以被视作是云端版的关系型或非关系型数据库，主要被应用于电子商务系统、银行、证券等行业。

云存储的优势首先在于其在扩大了数据存储能力的同时，规避了高昂的基础设施相关成本（有时这些成本也同时是本可避免的浪费，如分布式系统中可能存在的资源重复设置）。由于云存储具备弹性，企业在使用超出部分时可通过额外的付费来进行拓展，这使企业从如何预配置存储的艰难抉择中解放出来。同时不再需要花费时间与精力购买、维护基础设施，极大地减少了相应成本。其次，云存储技术极大地方便了对于数据的管理。成熟的云计算服务提供商往往提供用户友好型的使用界面，企业可以轻松地获取、操纵并分析这些数据。最后，云存储技术能够在一定程度上保证业务连续性：云计算服务提供商们拥有的数据中心往往具备更强的安全性并提供实时的检测，从而保证用户能从意外的操作或应用程序故障中恢复原有的数据库。

3.4　数据库管理

3.4.1　数据表

数据表的概念我们在介绍 SQL 技术时曾经提及过：数据表由行（row）和列（column）组成，是一个二维的网格结构，每一列都是一个字段。字段由字段名称和字段的数据类型以及一些约束条件组成，表中至少要有一列，可以有多行或 0 行，表名要唯一。

从更严格的层面来说，基于 SQL 的数据表往往包括如下元素：**实体**，客观存在并可相互区别的事物（可被看作数据表自身）；**属性**，包含属性名称和属性类别（可被看作表格中的一列）；**主键**，一类特殊的属性，用于界定表中记录的独特性，每一个体的主键都是独一无二的（如我们的身份证号）；**外键**，不同实体之间对于实体的引用，用于识别实体之间的联系，即不同数据表中通过某一共有列形成的联系，外键可以是不唯一的；**关系**，由外键确定的不同实体之间的相互连接；**基数**，实体与实体之间的关系中，

某一个体可能重复出现的次数。按照基数的不同，实体之间的关系可以是一对一、一对多或多对多的。

下面将以具体的例子来进行说明。以某大学课程数据库为例，数据库中包含着下列的 3 张表：教师信息表（表 3-3）、课程信息表（表 3-4）、教师课程表（表 3-5）。对应上文的定义：实体即为每一张表格；属性即为表格中的列，如课程信息表中的授课地点；主键即为每张表中 ID 性质的部分，如教师信息表中的教师工号；外键则是不同表之间可以连接的部分，如教师信息表和教师课程表中共有的教师工号一栏，课程信息表和教师课程表中共有的课程编号一栏；不同表之间通过共有列可以连接起来形成关系，其中教师信息表与教师课程表通过教师工号所建立的关系就是一对一的，而教师信息表与课程信息表所建立起的关系就是一对多的。

表 3-3　教师信息表

教师信息表	
教师工号	所属学院
J01	经济与管理学院
F02	法学院
S03	数学与统计学院

表 3-4　课程信息表

课程信息表		
课程编号	授课教师工号	授课地点
210720154	J01	A 教学楼
210720155	J01	B 教学楼
210720156	F02	A 教学楼
210720157	S03	A 教学楼
210720158	S03	B 教学楼
210720159	S03	C 教学楼

表 3-5　教师课程表

教师课程表		
教师 ID	所授课程	授课时间
J01	210720154，210720155	周二、周五
F02	210720156	周一、周四
S03	210720157，210720158，210720159	周一、周二、周五

基于数据表的管理方法在处理结构化数据时有着一些显著的优势。一方面，数据表在标准化方面拥有着巨大的优势；另一方面，基于 SQL 的数据表管理的逻辑是基于现实业务的逻辑的，可以比较直接地展现整个业务流程。但严格的表格结构显然不适用于多样的数据形式。

3.4.2　Hive 数据仓库

在介绍 Hive 数据仓库之前，我们先明确数据仓库的概念。数据仓库不同于数据库：数据仓库本身不生产任何数据而是从数据库或文件系统中拿数据并保存，数据仓库本身也不消费任何数据——数据分析的结果不是给自身使用而是用于驱动现实世界的决策。更加直接地说，设计数据仓库的目的不是**捕获数据**而是**分析数据**。举个例子，客户在银行的每笔交易都会写入数据库，可以理解成用数据库记账。数据仓库是分析系统的数据平台，它从事务系统中获得数据，通过分析、决策交易额、存款等维度来决定需要增加自动柜员机（automated teller machine，ATM）的地理位置。

通过数据表，我们可以清晰地整理业务流程中产生的数据，形成数据库并从中提取有价值的、能够指导实践的信息。那么，为什么我们需要一个专门用于分析的数据仓库呢？我们确实可以在基于 SQL 的数据库中直接对数据进行分析，但这么做一方面很容易对本地数据库或分布式系统造成极大的负荷，对原始数据的存储产生威胁；另一方面，数据库的数据往往是会更新的（周期为 1 周到 1 个月不等），因此无法对历史的数据进行管理和分析。由此数据仓库就诞生了。

数据仓库往往是分层的，其中的数据来自不同的数据源（如自己的服务器、第三方购买、抓取等）。数据自下而上流入数据仓库后，供上层使用。分层的一个目的在于通过大量的预处理（如筛选）来提升效率，因此数据仓库会存在大量的冗余数据。另一个目的是简化数据清洗过程：通过分层，可以将原来一步的工作分解成多个步骤，相当于将一个复杂的工作分解成多个简单的工作。这一操作使得每一层的逻辑都相对简单，出现问题时，局部调试对应步骤即可。不同公司数据仓库的分层形式不同。由于数据仓库不产生数据，也不消费数据，所以很自然地将数据仓库按照输入、存储、输出分为 3 层：操作数据存储层（operation data store，ODS）又叫源数据层、临时数据层，主要负责临时存放从数据源中提取的数据，作为数据仓库的输入；数据仓库层（data warehouse，DW）主要负责对 ODS 层提供的数据进行加工与整合；数据应用层（data application，DA）则面向业务为数据分析定制的数据。

Hive 就是一个在操作层面运用 SQL 语句，然后在应用内转换为另一种功能更多样、更复杂的代码（MapReduce）并执行的数据仓库。Hive 起源于 Facebook，它在官方网站上将自己定义为一个分布式的、具有容错性的数据仓库系统，为用户提供大规模分析的平台，并且允许 SQL 语句方便地读取、写入和管理分布式系统中的 PB 量级数据。这里需要特别注意，Hive 是一种数据仓库而非单纯的数据库。基于 Hive 的数据仓库管理的优点在于其在编程上的简易与较高的容量：一方面 Hive 的操作全部采用 SQL 语法，对于技术人员来说更容易上手，也更容易根据需求做出相应的分析；另一方面，在数据存储上，它能够存储很大的数据集，并且对数据完整性和格式的要求不严格。Hive 的主要缺点在于它的高延迟。Hive 操作默认基于 MapRreduce 引擎，而 MapReduce 引擎与其他的引擎相比，其特点就是慢、延迟高、不适合交互式查询。而 Hive 因为底层要转换成 MapRreduce，所以也有着高延迟的缺点。

3.4.3 访问权限

数据权限是指对系统用户进行数据资源可见性的控制，只有符合条件的用户才能看到该条件下对应的数据资源。比如，一个跨国企业的大中华区的销售人员是不能看见其他海外分区的客户信息的。设置数据权限的意义在于保护数据安全以及方便管理。如果一家企业的运营后台将所有可以查到的客户信息不加筛选都给到公司的每一个成员，并赋予他们修改、增删的权利，那么任何一个不慎的错误操作都可能会给企业造成巨大损失甚至法律风险。

设置数据相关的权限是进行数据管理中的一个重要环节，而梳理权限结构就是设计权限体系的第一步。数据的权限可以细分为数据查看权限、数据修改权限、数据读取权限等，对应到系统设计中有页面权限、菜单权限、按钮权限等。数据工作者需要根据实际情况梳理数据权限的结构。接下来的一步是思考怎么把权限分配给数据使用者。在数据的使用者较少的情况下，我们自然可以直接分配权限。但是随着用户数量的增长，每一个用户都需要单独地去分配权限，非常浪费数据库管理人员的时间和精力，并且用户和权限杂乱的对应关系会给后期带来巨大的维护成本。常见的权限管理模型包括自主访问控制（discretionary access control，DAC）、强制访问控制（mandatory access control，MAC）、基于角色访问控制（role-based access control，RBAC）和基于属性访问控制（attribute-based access control，ABAC）。

在使用 DAC 模型时，系统允许文件或资源的所有者自行决定谁可以访问它们。这种模型基于主体—客体—权限（subject-object-permission）模型，其中主体可以是用户或进程，客体可以是文件、文件夹或其他资源，权限则指主体可以对客体执行的操作。在 DAC 模型中，每个资源都有一个所有者，所有者可以授权其他用户或组访问资源，并指定所授予的权限。这种设计最常见的应用就是文件系统的权限设计，只有拥有权限的用户才可以对文件进行操作（读取、写入和修改等）。在 MAC 模型的设计中，每一个对象都有一些权限标识，每个用户同样也会有一些权限标识，而用户能否对该对象进行操作取决于双方权限标识的关系，这个关系的判断通常是由系统硬性限制的。比如，在特工相关的影视作品中我们经常能看到特工在查询机密文件时，屏幕提示“无法访问，需要一级安全许可”。这个例子中，文件上就有“一级安全许可”的权限标识，而用户并不具有。MAC 模型非常适合安全需求极度高的机构，但对于类似商业服务系统，则因为不够灵活而不能适用。

RBAC 模型可以解决权限系统中对应关系复杂的问题并设计出更贴合实际的数据权限系统。虽然现实中我们经常遇到这种用户和权限关系复杂的情况，但其实很多用户所需要的权限其实是一致或有部分重叠的，因此我们可以借助第三方媒介，把需要相同的权限都分配给这个媒介，然后用户和媒介关联起来，用户就拥有了媒介的权限了。这就是经典的 RBAC 模型，其中媒介就是我们通常所说的角色。有了角色之后就可以把权限分配给角色，需要相同权限的用户和角色对应起来即可。角色与权限的联系也是自由的。经典的 RBAC 模型中角色起到了桥梁的作用，连接了用户和权限：可以是一对多，也可以是多对一，这样用户就拥有了多个角色的多个权限。同时因为有角色作为媒

介，大大降低了错综复杂的交互关系。比如，一家有上万人的公司，角色可能只需要几百个就足够管理好所有权限了。例如，假设现在我们以老师和学生作为角色，以班级为单位建立了一个关于教学的数据库，那么我们就可以向班上的所有学生分配“学生”这一角色，向语文老师、数学老师、英语老师们分配“老师”这一角色，然后定义“学生”角色的权限为只能查看教学数据，定义“老师”角色的权限为查看、修改数据。

不同于常见的将用户通过某种方式关联到权限的方式，ABAC 根据主体（如用户）、客体（如资源）和环境（如时间和地点）的属性来控制访问。在 ABAC 模型中，访问策略规则基于多个属性组合，这些属性包括主体属性、客体属性、环境属性和操作属性。该模型具备着很强的灵活性：可以根据实际需要进行配置和管理，并且可以定义复杂的策略规则来控制访问，灵活地调整和修改这些规则以适应变化的需求。ABAC 模型的缺点，也是其没有被广泛推广的原因，主要是其在技术上和管理上的复杂性。但随着企业的业务流程关联越来越紧密，营销实践越来越复杂，将来 ABAC 有可能取代 RBAC 成为主流的权限管理模型。

在构建了权限管理模型之后，我们需要将其转化为数据库或数据仓库中实实在在存在的限制。无论是基于 SQL 的数据库还是基于 Hive 的数据仓库，都具备较为简单易操作的权限设置功能，数据工作者可以根据已搭建的数据权限管理模型去细化权限系统。

3.4.4　数据生命周期

数据生命周期是一个相当宏观的概念：从创建和初始存储，到使用于业务和日常维护，到最终过时被删除，是某个集合的数据从产生或获取到销毁的过程，其长度由数据自身的价值所决定。从整个数据生命周期的视角出发来管理数据是十分有必要的。随着数据爆发式的增长，越来越多的企业认识到数据的重要性，把数据当作一种珍贵的资产。但其实数据并不等价于企业的数据资产——数据必须也只有以合理、易用、安全和易于理解的方式组织起来，能为业务注入有效的价值，才能成为企业的数据资产。所以数据变成数据资产的前提是包含数据标准管理、数据质量管理、数据安全管理和持续产生数据价值管理在内的从数据产生到销毁的数据全生命周期管理体系。

数据全生命周期管理体系下，数据生命周期的每个阶段都由一组策略管理，目的是在生命周期的每个阶段最大限度地发挥数据的价值。具体的步骤如图 3-2 所示。

第一阶段是数据创建。在此阶段，企业主要需要关注数据的质量。新的数据生命周期始于数据收集，但数据来源过于丰富，如 Web 和移动应用、物联网设备、表单、调研等。虽然数据可能是通过各种不同的方式生成的，但并非所有可用数据对企业的成功都是必不可少的。而冗余的数据会带来额外的存储和管理成本。因此，必须始终根据数据质量及其与企业的相关性，评估是否需要纳入、整合新数据。具体来说，在这一阶段，企业需要解决从何处收集信息、收集何种信息、如何收集信息以及数据库或数据仓库的初步建立等问题，同时需要特别注意日渐瞩目的个人信息保护问题。

第二阶段是数据存储。数据所采用的结构可能各不相同，这会对企业使用的数据存储类型产生影响。结构化数据一般使用关系型数据库，而非结构化数据（如文本、图片、

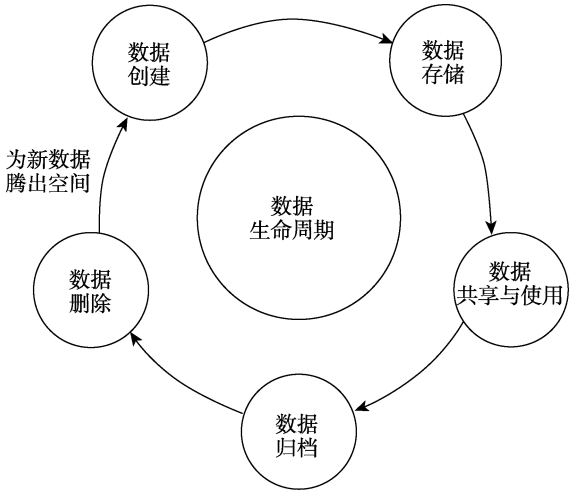

图 3-2　数据生命周期管理

音频和视频）通常使用 NoSQL（即非关系型）数据库。确定适用于数据集的存储类型后，就可以评估基础架构是否存在安全漏洞，以及是否可对数据进行各种不同类型的处理，如数据加密和数据转换，以保护企业免受恶意行为实施者的威胁。这一步骤还可确保敏感数据遵守隐私和政府政策的要求，帮助企业避免违反此类法规而导致的巨额罚款或可能导致的相关诉讼。数据存储还聚焦于数据冗余。任何存储的数据都应该有一份副本作为备份，以防出现数据删除或数据损坏等情况，并在网络安全问题日趋严重的今天防止对数据的意外更改以及包括恶意软件攻击在内的蓄意破坏。

第三阶段是数据共享与使用。在此阶段，企业需要定义数据的使用者及数据的用途。确定数据可用后，就可以进行一系列分析，包括基本的探索性数据分析、数据可视化以及更高级的数据挖掘和机器学习方法。所有这些方法都在业务决策以及与各利益相关者的沟通中发挥重要作用。此外，数据使用并不一定限于使用内部。例如，外部服务提供商可出于营销分析和广告等目的使用数据。内部使用包括日常业务流程和工作流程，如仪表板和演示等。

第四阶段是数据归档。经过一段时间的使用后，一些数据对于具体的业务可能不再有用。但是企业必须保留不经常访问的组织数据的副本，以用于满足可能的诉讼和调查要求。如果需要，可将归档的数据恢复到活动的生产环境中。在这一阶段，企业应该明确定义何时归档数据、归档到何处以及归档多长时间，并确定数据经历归档过程以确保冗余。

第五阶段也是最后一个阶段，是数据删除。在数据生命周期的最终阶段，数据被从记录中清除并安全销毁。企业将删除不再需要的数据，以便为新的仍具有价值的数据腾出更多存储空间。在此阶段，当数据超过要求的保留期或不再对组织具有有意义的用途时，将被从归档中删除。

数据生命周期管理具有几个重要的优点，包括流程改进、控制成本、数据易用性、合规与治理。首先，数据在推动组织的战略计划方面发挥着至关重要的作用。数据生命周期管理能够在数据的整个生命周期中保证其质量，从而帮助改进流程和提高效率。出

色的数据生命周期管理战略可确保供用户使用的数据准确而可靠，帮助企业最大程度发挥数据的价值。其次，数据生命周期管理流程在数据生命周期的每个阶段实现其价值。一旦数据对生产环境不再有用，组织可利用一系列解决方案以降低成本，包括数据备份、复制和归档。例如，可将数据转移到本地、云或网络连接存储中低成本的存储位置。再次，借助数据生命周期管理战略，数据工作者可以制定策略和规程，确保以一致的方式标记所有元数据，以便在需要时提高数据的可访问性。建立可执行的治理策略，确保数据在其保留期内一直发挥价值。清洁有用的数据的可用性有助于提高企业流程的敏捷性和效率。最后是合规与治理方面。每个行业领域都有自身的数据保留规则，因此强有力的数据生命周期管理战略可以帮助企业保持合规合法。此外，数据生命周期管理使企业能够以更高的效率和安全性来处理数据，同时确保遵守有关个人数据和组织记录的数据隐私法律。

对于企业来说，数据生命周期管理并不是某一部门独立能够完成的任务，它涉及了本章几乎所有的内容，是一项需要整个组织上下通力协作才能完成的工作。数据生命周期管理对应的组织架构可以分为决策层、管理层和最终的执行层，营销业务只不过是执行层中的一小部分。因此在数据全生命周期管理体系下，我们不能孤立地看待营销业务，而应该以系统的视角来展开具体的实践。

3.5　案　　例

长安汽车是中国知名汽车制造企业，中国汽车品牌行业领跑者之一。长安汽车从 2001 年开始信息化建设，在 2010 年开始进入数字化转型阶段，在 2019 年开始进入数字化重塑阶段，完成了传统制造业在大数据时代的华丽转型。

2000 年是长安汽车信息化建设的一个分水岭。因为业务战略的转变，长安汽车此前建立的分布在各部门、各领域的信息系统已经无法满足新的业务发展的需要。为了长远发展，长安汽车随后成立了专门的信息公司，开始全面推进信息化建设。长安汽车通过统一平台、统一数据、统一运营的数字化管理手段，将数据治理的成果应用于生产实践，实现基础管理标准化、业务数据财务化、经营成果指标化。长安汽车数字化转型的另一个成果是整合了企业数据，建立起了企业自己的大数据平台。长安汽车整合企业内部核心业务系统，通过融合互联网数据、统一客户数据并集成产品数据，构建了集数据管理、分析、应用于一体的“长安−数据驱动管理”（Chang An-data drive management，CA-DDM）大数据平台（以下简称“大数据平台”），服务公司经营，以实现“现状可见、问题可查、风险可辨、未来可测”，促进数据的全面、真实、透明、共享。

大数据平台的总注册用户人数达 7200 余人，月均访问量近 30 万人次，广泛应用于 12 个领域，推动了 60 多项管理制度的改善。这一大数据平台的功能在于：首先，完善数据门户，随时随地获取可信任的数据，实现平台的数据化、系统化、透明化管理，面向全员开放，及时掌握各领域及各部门的大数据运营状态，以及共同建立和持续改善数据文化氛围；其次，构建数据分析工作台，让业务人员可在线进行数据分析，效益提升了 40%；最后，建立数据实验室，研究数据分析模型等，将销售预测、客户洞察、市场

洞察等核心模型投入业务应用中。

在完善、全面、系统的数据管理框架下，长安汽车在营销领域突破了传统业务模式，促进了业务创新。例如：深度解析制造过程，提高了生产效率；实现了多维用户画像，开展用户个性化服务；发现潜在人群和需求，实现精准营销，深入挖掘了潜在的客户资源。长安汽车的数字化转型在建立总体数据管理体系，打通不同业务方面为营销实践者树立了良好的表率，也为传统制造业在大数据时代的转型提供了深刻洞见。

3.6 本章小结及习题

3.6.1 本章小结

（1）消费者数据的来源可以是企业自行调研、在线数据和购买或申请的来自第三方的数据。

（2）企业获取消费者数据的传统方法包括观察法、调查法和实验法。随着网上问卷平台的兴起，企业越来越多地选择网上营销调研的方式。

（3）企业的在线数据获取渠道包括可以来自在其他平台或 App 上广告的推送，自有的网站或应用，事件监测（埋点）和公众号或小程序端等。

（4）数据可以被分类为结构化数据和非结构化数据，但最终都需要被转化为用于分析、能够提升业务的行为数据。

（5）非结构化数据包含文本数据、视频流数据等形式不定的数据，往往具备着更高的价值。

（6）大数据的五个基本特征包括体量大、多样化、速度快、价值密度低、真实性（volume、variety、velocity、value、veracity）。

（7）NoSQL 及分布式系统不仅在面对海量且多样的数据时更加游刃有余，而且具备更高的敏捷性。

（8）企业应该从整个数据生命周期的角度对数据进行管理，并通过数据库或者数据仓库实现，其中一个重要的要素是数据权限。

（9）设置数据相关权限是数据管理的关键一步。管理者可以根据需求选择 DAC、MAC、RBAC 和 ABAC 其中的一种。

（10）数据生命周期管理是一种理念，着眼于数据从产生到消亡的历程，从整个组织出发，高屋建瓴地对数据进行系统的管理。其优势在于流程改进、控制成本、数据易用性、合规与治理。

3.6.2 习题

马自达汽车物流公司负责马自达汽车和零部件在欧洲的分销，其传统的仓库管理系统缺乏运输管理模块，使马自达公司无法了解实时的产品发货情况，也无法对承运人预

订、发票和账单等数据进行有效的管理。这使得马自达公司无法确保按时交货，不仅危及客户满意度，还可能使竞争对手从马自达手中夺走市场份额。随着汽车行业从“以产品为中心”到“以客户为中心”，马自达公司希望对其旗下的汽车物流业务进行数字化的改造，改善用户体验，并更多地将资源从日常事务转移到客户服务上。

问题：

（1）马自达汽车物流公司需要建立的运输管理模块主要产生什么类型的数据？

（2）马自达汽车物流公司需要的数据存储和管理技术是哪一种？

（3）从数据生命周期的视角出发，马自达汽车物流公司在各个数据生命阶段的策略应该是什么样的？

参考文献

[1] CHARLESWORTH A. Digital marketing: a practical approach[M]. Abingdon: Routledge, 2014.

[2] WIRTH N. Hello marketing, what can artificial intelligence help you with?[J]. International Journal of Market Research, 2018, 60(5): 435-438.

[3] VENKATESAN R, LECINSKI J. The AI marketing canvas: a five-stage road map to implementing artificial intelligence in marketing[M]. Palo Alto: Stanford University Press, 2021.

[4] 曾凡涛，熊元斌. 试论数据挖掘技术在旅游营销中的应用[J]. 旅游科学，2002(4): 13-16.

[5] 惠琳. 大数据时代本土零售业精确营销探讨：基于数据挖掘的角度[J]. 商业时代，2014(4)：58-59.

[6] 贾建民，杨扬，钟宇豪. 大数据营销的“时空关” [J]. 营销科学学报，2021，1(1)：97-113.

[7] 万红玲. 大数据时代下的精准营销[J]. 新闻传播，2014(1)：71-71.

[8] 张俊杰，杨利. 基于大数据视角的营销组合理论变革与创新[J]. 商业经济研究，2015(6)：54-55.

即测即练

自学自测　扫描此码